UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Gesture recognition using deep learning in virtual reality environments

A Degree Thesis

Submitted to the Faculty of the

Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona

Universitat Politècnica de Catalunya

by

José Antonio Forcada Sans

In partial fulfilment

of the requirements for the degree in

AUDIOVISUAL ENGINEERING

Advisor: Javier Ruiz Hidalgo

Barcelona, May 2018

Abstract

Contactless systems have taken on a notable presence in the evolution of human-computer communication in recent years.

Methods based on gestural recognition have been one of the most investigated.

With the emergence of deep learning models and the increase of hardware possibilities, they have allowed us to work with increasingly larger databases and with larger and more complex architectures. Being able to improve previous results.

This document introduces the implementation of a Gestual Recognition model in Deep Learining in virtual environments using color and depth images (RGB-D).

To achieve this objective, a model inspired by the R3DCNN model presented in the work of P. Molchanov [2] trained in the Chalearn database and in the Telepresence Dataset database created by the student José Famadas in his Final Degree Thesis [1].

With this model, precision values of 67% have been obtained for the Chalearn database and 27% for the Telepresence database. This makes us think that we have been able to improve previous results of other projects, although it has not been possible to deal with the problems with the Telepresence database.

Resum

Els sistemes que no requereixen d'un contacte directe han agafat una presència notable en la evolució de la comunicació entre les persones i les màquines en els darrers anys.

Els mètodes basats en el reconeixement gestual han estat un dels més investigats.

Amb la irrupció dels models d'aprenentatge profund (Deep Learning) així com l'augment de les possibilitats de hardware han permès treballar amb bases de dades cada cop més grans i amb arquitectures més grans i complexes. Podent millorar els resultats anteriors.

En aquest document s'introdueix la implementació d'un model de Reconeixement Gestual en Deep Learining destinat a entorns virtuals fent servir imatges en color i profunditat (RGB-D).

Per assolir aquest objectiu es fa servir un model inspirat en el model R3DCNN presentat en el treball de P. Molchanov [2] entrenat en la base de dades de Chalearn i en la base de dades Telepresence Dataset creada per l'estudiant Josep Famadas en el seu treball de final de grau [1].

Amb aquest model s'ha obtingut uns valors de Accuracy del 67% per la base de dades de Chalearn i d'un 27% per la base de dades de Telepresence. El qual ens fa pensar que s'ha aconseguit millorar els resultats previs d'altres projectes malgrat que no s'ha pogut fer front als problemes amb la base de dades de Telepresence.

Resumen

Los sistemas que no requieren de un contacto directo han cogido una presencia notable en la evolución de la comunicación entre las personas y las máquinas en los últimos años.

Los métodos basados en el reconocimiento gestual han sido uno de los más investigados.

Con la irrupción de los modelos de aprendizaje profundo (Deep Learning) así como el aumento de las posibilidades de hardware han permitido trabajar con bases de datos cada vez más grandes y con arquitecturas más grandes y complejas. Pudiendo mejorar los resultados anteriores.

En este documento se introduce la implementación de un modelo de Reconocimiento Gestual en Deep Learining destinado a entornos virtuales utilizando imágenes en color y profundidad (RGB-D).

Para alcanzar este objetivo se utiliza un modelo inspirado en el modelo R3DCNN presentado en el trabajo de P. Molchanov [2] entrenado en la base de datos de Chalearn y en la base de datos Telepresence Dataset creada por el estudiante José Famadas en su trabajo de fin de grado [1].

Con este modelo se han obtenido unos valores de precisión del 67% para la base de datos de Chalearn y de un 27% para la base de datos de Telepresence. El que nos hace pensar que se ha conseguido mejorar los resultados previos de otros proyectos, aunque no se ha podido hacer frente a los problemas con la base de datos de Telepresence.

Agraïments

Primer de tot agrair al meu tutor Javier Ruiz Hidalgo per haver-me ofert la possibilitat de realitzar aquest treball, així com la possibilitat d'introduir-me en la recerca en el món del Deep Learning, i també pel suport en la realització de les diferents tasques portades a terme.

Per altra banda agrair al departament de Imatge de la UPC, no només pels recursos oferts sinó també pel suport rebut. En el mateix sentit agrair també al grup d'investigació de Telepresència de la UPC així com també al suport tècnic rebut per part de l'Albert Gil qui m'ha ajudat a tirar part de la tasca de implementació endavant.

Finalment, i no menys important, agrair la feina per part dels diferents professors/es de l'ETSETB en la obtenció dels coneixements rebuts durant aquests anys. Coneixements que m'han permès afrontar aquest repte.

Revision history and approval record

Revision	Date	Purpose
0	15/04/2018	Document creation
1		Document revision

DOCUMENT DISTRIBUTION LIST

Name	e-mail
José Antonio Forcada Sans	jose.antonio.forcada@alu-etsetb.upc.edu
Javier Ruiz Hidalgo	

Written by:		Reviewed and approved by:	
Date	15/04/2018	Date	
Name	José Antonio Forcada Sans	Name	Javier Ruiz Hidalgo
Position	Project Author	Position	Project Supervisor

Índex

Llistat de figures

Llistat de Taules:

1. Introducció

La evolució de la comunicació entre les persones i les màquines ha estat un dels punts claus en el progrés i la integració de les tecnologies TIC en al vida de les persones. Així doncs si fins fa uns anys fins la actualitat aquesta comunicació es feia a través d'interfícies preparades per a aquest objectiu, la evolució de les tecnologia immersives, en concret la Realitat Virtual, permeten que aquestes interfícies deixin de ser necessàries doncs la comunicació entre les persones i les màquines passa a ser directe. Aquesta comunicació pot es pot donar en diverses formes com la verbal (reconeixement de veu i ordres), així com el contacte directe amb l'aparell (sistemes tàctils). Dins d'aquestes formes de comunicació, la gestual és una de les més investigades. Dins d'aquesta investigació la recerca de mètodes i arquitectures que permetin el reconeixement de gestos i accions per tal de poder ser interpretades per les màquines ha estat un dels punts claus.

1.1. Antecedents del projecte

Aquest projecte forma part del projecte de Telepresència portat a terme per diferents estudiants de Màster com estudiants de grau dins Departament d'Imatge de la Universitat Politècnica de Catalunya i amb la supervisió dels professors Javier Ruiz Hidalgo, Josep R. Casas i el suport tècnic de l'Albert Gil Moreno.

Aquest projecte és pot entendre com una continuació del treball de final de grau de l'estudiant Josep Famadas [1]. En ell implementava una xarxa neuronal 3D CNN amb l'objectiu de reconèixer gestos a partir d'una base de dades que ell mateix creà.

Els seus resultats no van ser dolents en quan a la implementació però no s'obtingueren bons resultats al aplicar el model sobre la base de dades creada.

L'objectiu d'aquest projecte és seguir doncs en la investigació portada a terme per aquest estudiant i el grup de recerca tot aplicant nous mètodes i noves implementacions.

1.2. Declaració de pròsits

Aquest projecte s'ha dut a terme dins el Departament de Teoria de la Senyal i Comunicacions de l'Escola Tècnica Superior de Telecomunicació de Barcelona.

El projecte consistirà en la implementació d'un model C3D RNN per al reconeixement gestual seguint els passos del treball de en Josep Famadas [1], com de Molchanov i el grup de recerca de NVIDIA [2], així com fent servir part de les implementacions de l'estudiant Alberto Montes en el seu treball de fi de grau [4].

En resum, els passos que s'han seguit han sigut els següents:

1) Prendre un model ja entrenat i realitzar un re-entrenament (finetunning) amb la base de dades de ChalLearn 2016. Usant el framework de Caffe (Universitat de Barkley).[5]
2) Si els resultats són positius aplicar el mateix procediment per a la base de dades de Telepresència.
3) Extreure característiques mitjançant els models entrenats i fent servir una modificació de la implementació del treball de l'Albert Montes amb la plataforma Keras, aplicar una capa recursiva LSTM al nostre model per tal de explotar la dimensió temporal amb més profunditat. Això aplicat a les dues bases de dades.

1.3. Requeriments i especificacions

Els requeriments d'aquest projectes són en definitiva:

- Re-entrenar sobre les dues bases de dades(ChalLearn 2016 i Telepreence) un model C3D prèviament entrenat amb la base de dades de l'UCF101.
- Un cop analitzats els resultats obtinguts aplicarem l'esquema proposat per el treball de l'Albert Montes [4], on fa servir el model ja re-entrenat per extreure característiques i entrenar un model afegint una capa recursiva. Aquest segon pas es realitza amb la plataforma Keras.

Com hem dit abans, el primer pas es fa fent servir un implementació per C3D [3] sobre la plataforma de Caffe [5] de la Universitat de Berkaley. Aquesta plataforma es usada en investigació acadèmica i ofereix moltes possibilitats, així com facilitats doncs només ens cal preparar arxius de configuració evitant la tasca de programació des de zero.

En el segon punt s'ha fet servir Keras [6]. La raó d'això és el fet que Caffe no té una implementació gaire flexible de capes recursives LSTM. Doncs estan implementades amb l'objectiu de ser aplicades al reconeixement de llenguatge i no tant en la classificació de vídeo. Vist que Caffe no ens ofereix aquesta possibilitat agafant l'exemple de l'Albert Montes s'ha procedit a passar a fer servir la plataforma de Keras. Keras és una llibreria de models que en aquest cas utilitzant la plataforma de Theano ens permet implementar capes Recursives de manera més àgil i flexible. El fet que estigui desenvolupada en Python ens permet poder adaptar-la a les necessitats del projecte.

1.4. Mètodes i procediments

Com s'ha descrit anteriorment el primer pas ha estat realitzar un re-entrenament sobre cadascuna de les dues Bases de Dades d'un model ja prèviament entrenat sobre la base de dades UCF101.

Exactament el que estem fent és agafar un model i els seus respectius weights (pesos) entrenats en una base de dades i els re-entrenem per tal de adaptar-los a la base de dades que ens interessa. Aquest re-entrenament es realitza agafant clips de vídeo, i segmentant-los en seqüencies de 16 frames formant sub-clips.

Un cop fet aquest primer pas procedim mitjançant les llibreries de Keras i la plataforma Theano a realitzar l'entrenament en el model complet.

Primer extraurem les característiques de la últim capa Fully Connected del model C3D primari, això ens donarà un total de 4096 característiques per cadascun dels sub-clip de vídeo.

Un cop realitzat aquest pas agafarem aquestes característiques i les introduirem a un model amb una capa LSTM. Finalment per cadascun dels sub-clips de vídeo obtindrem una classificació, per lo que recorrerem a un capa SoftMax que ens permetrà assignar una classe a cadascun dels nostres vídeos.

L'esquema per tant és el següent.

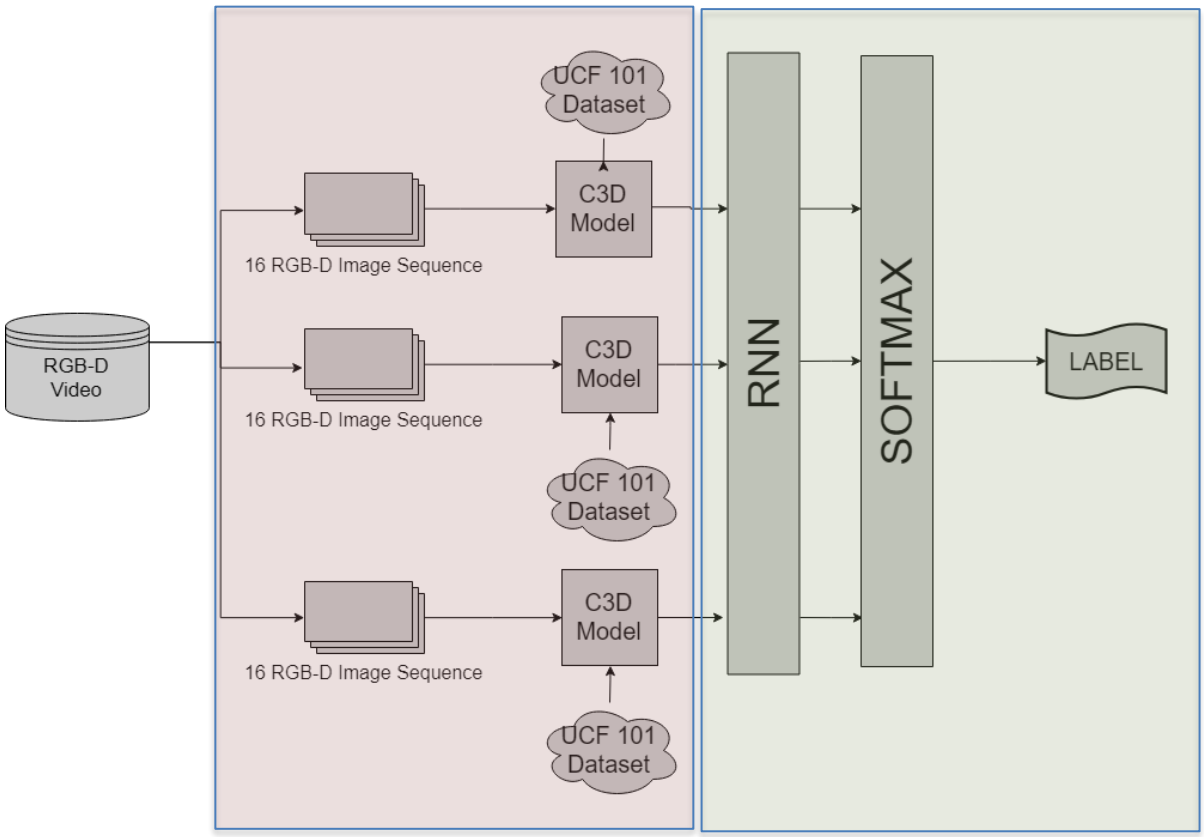

Figure 1.1: *Esquema C3DRNN*

1.5. Pla de treball

El pla de treball principal no ha estat complert del tot doncs la pròrroga ha allargat els temps en alguns Work Package's.

WP#	Task#	Short title	Date (week)
1	1	Research of a previous Project as guideline	27/10
2	2	Setting resources for the guideline project.	13/11
3	3	DL model implementation and training	14/12
4	4	VR environment adaptation	20/01
5	5	Full model training and testing	24/03

Table 1.1: *Work Packages table.*

1.6. Diagrama de Gantt

	Modo de	Nombre de tarea	Duración	Comienzo	Fin
1		**Project Proposal and Workplan**	**6 días**	**vie 29/9/17**	**vie 6/10/17**
2		1st Meeting	1 día	vie 29/9/17	vie 29/9/17
3		2n Meeting	1 día	vie 6/10/17	vie 6/10/17
4		**Setting a previous Project as guideline**	**18 días**	**sáb 7/10/17**	**vie 27/10/17**
5		Analysis of the state of the art	2 días	sáb 7/10/17	lun 9/10/17
6		Find an appropriate	16 días	mar 10/10/17	vie 27/10/17
7		3rd Meeting	1 día	vie 13/10/17	vie 13/10/17
8		**Setting resources for the guideline project.**	**12 días**	**sáb 28/10/17**	**lun 13/11/17**
9		Analysis of the different candidates	3 días	sáb 14/10/17	mar 17/10/17
10		Obtention of resources	20 días	mié 18/10/17	lun 13/11/17
11		4th Meeting	1 día	vie 20/10/17	vie 20/10/17
12		**Implementation of the guideline DL environment**	**27 días**	**mar 14/11/17**	**jue 14/12/17**
13		Implementation of the Guideline Project DL environment.	22 días	mar 14/11/17	vie 8/12/17
14		4th Meeting	1 día?	vie 27/10/17	vie 27/10/17
15		5th Meeting	1 día?	vie 3/11/17	vie 3/11/17
16		6th Meeting	1 día?	vie 10/11/17	vie 10/11/17
17		7th Meeting	1 día?	vie 17/11/17	vie 17/11/17
18		Training of the model.	3 días?	sáb 9/12/17	mar 12/12/17
19		8th Meeting	1 día?	vie 24/11/17	vie 24/11/17
20		Results analysis	2 días	mié 13/12/17	jue 14/12/17
21		**VR environments adaptation**	**32 días**	**vie 15/12/17**	**vie 26/1/18**
22		Training and adaptation of the model with an own DB.	27 días	lun 4/12/17	vie 5/1/18
23		9th Meeting	1 día?	vie 8/12/17	vie 8/12/17
24		10th Meeting	1 día?	vie 15/12/17	vie 15/12/17
25		11th Meeting	1 día?	vie 22/12/17	vie 22/12/17
26		Analysis of the results.	15 días	lun 8/1/18	vie 26/1/18
27		12th Meeting	1 día?	vie 5/1/18	vie 5/1/18
28		13th Meeting	1 día?	vie 12/1/18	vie 12/1/18
29		14th Meeting	1 día?	vie 19/1/18	vie 19/1/18
30		15th Meeting	1 día?	vie 26/1/18	vie 26/1/18

Figure 1.2: *Gantt Diagram*

2. Estat de l'art

Com ja s'ha esmentat prèviament el reconeixement gestual és una de les formes de interacció entre les màquines i les persones que més han estat investigades en les darreres dues dècades.

Amb la irrupció del Deep Learning els resultats per als experiments en aquests sentit han resultats ser molt bons fent que actualment l'state-of-the-art s'encamini en aquesta direcció en detriment dels mètodes "non-deep". Tot i això un dels reptes més importants en els quals s'ha de fer front és la explotació de la dimensió temporal en les seqüencies, doncs aquesta és clau en el reconeixement gestual, però augmenta considerablement la complexitat dels models.

Per abordar aquesta problemàtica s'han presentat diverses arquitectures i estratègies per abordar aquest problema. La que en els darrers any està agafant més protagonisme son les arquitectures basades en el modelatge temporal. Ja fa un anys es va investigar els models recurrents com RNN (Recursive Neural Networks) que pretenien abordar el problema però es trobaren en moltes dificultats en quan a càlcul. Anys més tard es començaren a estudiar les cel·les LSTM (Long Short Term Memory) en les RNN. Actualment les LSTM són claus en múltiples aplicacions com predicció textual, reconeixement de veu... I usades per grans corporacions com Google o Facebook.

2.1. Arquitectures

Com hem dit abans explotar les dependències temporals és clau per al reconeixement gestual fent servir Deep Learning. Per a afrontar aquest repte es presenten 3 grups amb diferents arquitectures:

- **Filtres 3D en les capes convolucionals:** L'ús de filtres de tres dimensions tant en les capes de Convolució com les de Pooling permeten explotar les dimensions espacials i la temporal. Un dels models més usats és el model C3D, que serà peça clau en el model presentat en aquest projecte.

  Model C3D.

  Aquest model presentat pels investigadors D. Tran, L. Bourdev, R. Fergus, L. Torresani, i M. Paluri en el seu treball [3], és un dels més usats actualment en el reconeixement gestual. Aquest model presenta 8 capes de convolució , 5 de Max-Pooling, i 2 Fully-Connected, seguides d'una capa de sortida Softmax. L'estructura es representada en la figura següent.

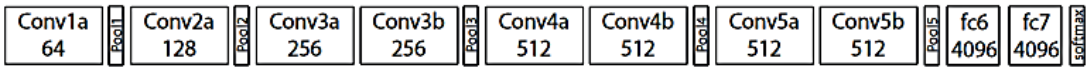

Figure 2.1: Esquema del model C3D

  Aquest model ha aconseguit bons resultats i és el que es farà servir en aquest projecte com a part del model complet.
  En el treball dels investigadors abans esmentats obtenen bons resultats en la base de dades **UCF101**, agafant com a mesura l'accuracy en el fase de test, obtingueren un valor de **82.3%** amb una sola xarxa i per vídeos RGB.

- **Extracció de característiques de moviment:** Aquestes arquitectures realitzen un pre-processament a les seqüencies d'imatges obtenint característiques temporals com els Optical Flow Maps. Aquestes característiques temporals sumades a les espacials permeten afrontar el repte de explotar la dimensió temporal en el reconeixement gestual. Ara bé, dintre del context del Deep Learning resulta poc eficient haver de realitzar un pre-processament doncs precisament el gran benefici del Deep Learning és evitar un processament previ de les dades per a l'extracció correcte de característiques. És per això que aquest mètode quedà descartat al començament d'aquest projecte. La raó és que el principal interès era aprofitar totes les possibilitats que ens ofereix el Deep Learning.

- **Models Deep Learning Temporals: RNN i LSTM**
  Com anteriorment s'ha esmentat una de les arquitectures més usades per explotar la dimensió temporal és l'ús de arquitectures recursives que ens permetin explotar la dimensió temporal en les seqüencies. Dintre d'aquestes capes les LSTM com hem dit abans són les més usades. Aquestes capes recursives són usades en combinació amb les xarxes presentades anteriorment. En aquest projecte la capa LSTM és usada en la capa Recursiva.

Model R3DCNN

És un model presentat per P. Molchanov i el grup de recerca de NVIDIA []. En el seu projecte es proposa un model que combina el model C3D afegint una capa recursiva a la part final. L'esquema presentat és el següent:

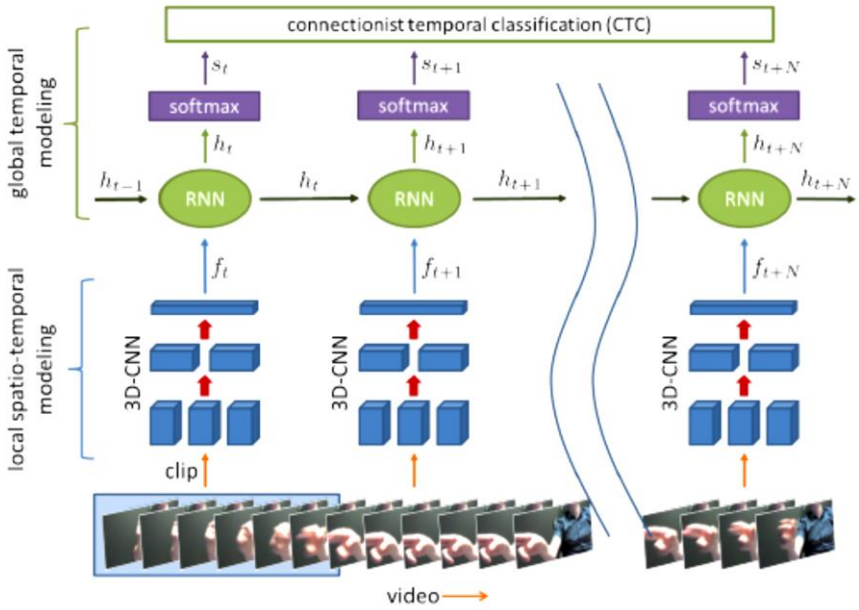

Figura 2.2: *Esquema R3DCNN de P.Molchanov[]*

Aquest model fa servir el model 3DCNN per extreure les característiques i hi afegeix una capa RNN per tal de d'aprofitar la informació seqüencial. Per calcular els paràmetres d'una seqüència en un instant necessitarem els paràmetres d'instants anteriors. Aquesta tècnica anomenada Back Propegation Trough Time (BPTT) és usada en reconeixement del llenguatge així com de veu. Ara vé en el cas que ens ocupa aquesta tècnica té dificultats a l'hora de ser aplicada per seqüències gran on es requereix fer ús de les dependències a llarg termini. Una

de les tècniques que es fa servir en aquest treball és recórrer a les capes LSTM que han demostrat solucionar aquests problemes.
Finalment en aquest treball s'afegeix una capa CTC (Connectionist temporal classification) per obtenir la funció de cost. Aquesta capa CTC és una funció de cost molt útil per a seqüències de vídeo no segmentades. En el projecte de P. Molchanov[] es treballa amb seqüències de vídeo llargues sense retallar per lo que té sentit aquest càlcul tenint en compte la possible classe "no gest". En el nostre cas treballarem amb bases de dades on totes les seqüències de gestos han estat segmentades.

2.2. Bases de dades

Un cop explicades les principals arquitectures i quines han sigut la base del nostre projecte, ens cal veure quines són les base de dades que s'usen en l'actualitat per al reconeixement gestual.

Com s'ha dit anteriorment una de les vies de investigació més importants és en la creació de base de dades mes grans que permetin un millor modelatge.

Pel reconeixement gestual les principals Datasets són:

- **ChaLearn:** La primera de 2014 es tracta d'un conjunt de 14000 Vídeos , amb 20 gestos del llenguatge de signes italià, realitzats per diferents persones. Es una de les base dades més grans i es realitzà un Challenge.
  En el treball de P. Molchanov entrenen el seu model R3DCNN amb aquesta base de dades.

  L'any 2016 es realitzà un altre challenge en el reconeixement gestual. Per aquest nou challenge la base de dades de Chalearn es formà per 47933 Vídeos RGB-D, de 249 gestos i múltiples usuaris.

  Aquesta segona base de dades serà amb la que realitzarem els nostres experiments del nostre model.

- **VIVA data set:** Dataset propi de NVIDIA i que per tant no és obert.

- **UCF-101:** Finalment i tot i que no és del tot per al reconeixement gestual, el data set de la Universitat Central de Florida és força gran i està pensada per al reconeixement d'accions i no tant de gestos. Està basada en 13320 clips RGB de vídeos de 101 accions humanes. El model C3D pre-entrenat que agafarem haurà estat entrenat amb aquesta base de dades.

- **UPC Telepresence Dataset:** Dataset creat per l'estudiant de la UPC Josep Famadas en el seu projecte final [1] basat en 270 videos RGB-D (Depth, profunditat capturada amb una càmera Kinect), amb 6 classes i amb 9 usuaris.

2.3 Llibreries i entorns de treball

Hi ha diverses frameworks usades per al desenvolupament de implementacions en Deep Learning. Les principals són:

Caffe: Software desenvolupat per Yangqing Jia amb la BVLC (Berkeley Vision and Learning Center). Està escrit en C++ i accepta també Python. És dels més utilitzats en l'àmbit de la recerca i un dels més potents. A més disposa de molts exemples

d'implementacions en diferents repositoris. Un dels principals inconvenients és que al treballar bàsicament amb arxius de configuració, es fa difícil poder afegir nous mètodes.

Tensorflow: Desenvolupat per Google i de codi obert, és una dels més utilitzats en la indústria i la recerca. Està escrit en C++ però accepta llenguatges com Python.

Theano: Llibreries escrites en Python de codi obert i en un primer moment desenvolupat per la Universitat de Montreal.

Keras: Llibreries per a Deep learning que fa servir altres plataformes com Theano o Tensorflow, aprofitant així els beneficis dels dos.

Finalment com a resum podem agafar la taula creada per Justin Johnson de la Univeristat de Standford [7].

	Caffe	Torch	Theano	TensorFlow
Language	C++, Python	Lua	Python	Python
Pretrained	Yes++	Yes++	Yes (Lasagne)	Inception
Parallel GPUs: Data	Yes	Yes	Yes	Yes
Parallel GPUs: Model	No	Yes	Experimental	Yes (best)
Readable Source Code	Yes (C++)	Yes	No	No
Good at RNN	No	Mediocre	Yes	Yes (best)
Higher-Level APIs	No	No	Keras	Keras and TFLearn

Table 2.3: *Justin Johnson Frameworks Comparaison 2016 cs231n Lecture 8*

2.4 Elecció de l'entorn de treball:

Com hem vist de les tres principals arquitectures presentades les que farem servir en aquest projecte són les basades en Convolucions en 3D (3DCNN) en concret el model C3D i les basades en RNN fent servir LSTM. Així doncs en aquest projecte farem servir una capa C3D per extreure característiques espai-temporal i finalment una capa recursiva LSTM al final per explotar la dimensió temporal a llarg termini.

Un cop vistes les base de dades principals s'escull les bases de dades de Chalearn y la de l'UPC Telepresence Dataset. La primera s'escull per que és la més gran que hi ha actualment i malgrat que el treball de referència de Molchanov agafa el Dataset de ChalLearn de 2014, el Dataset de 2016 té més classes i és molt més gran. Per altra banda el Dataset de la UPC, Telepresence Dataset, ens permetrà veure si es podem millorar els resultats obtinguts per l'estudiant Josep Famadas, qui realitzà aquest Dataset però no obtingué gaire bons resultats tot i aplicant un model 3DCNN (prenent de base un estudi anterior de Molchanov[2]) com és el cas d'aquest projecte.

Finalment observant els diferents frameworks possibles, s'escull una versió de Caffe com a implementació del model C3D i l'extracció de les característiques de les dues bases de dades i finalment agafem Keras per a la implementació del model complet.

3. Metodologia / desenvolupament del projecte:

Un cops establerts els passos a seguir el projecte s'inicia agafant la implementació de Caffe per C3D [3] Desenvolupada per Facebook. Veure Apèndix 1.

A més s'obtenen:

- Les dues bases de dades: ChaLearn i Telepresence. En el cas de la base de dades de Telepresence, al ser petita, caldrà fer una Validació Creuada (Cross Validation).
  S'ha de tenir en compte que la base de dades de Chalearn només podem disposar etiquetes per al set de traint, per lo que només podrem fer servir aquest en el nostre projecte. Per tant dels 47 933 vídeos del Dataset, nomes podrem fer servir 35 878 vídeos.
- També s'obté el model pre-entrenat en al data set UCF 101.

Es creen sub-clips de 16 frames per cada vídeo. Aquests vídeos són de 320x240 en el cas de la base de dades de Chalearn i de 512x424. Per a l'entrenament es realitza un retall per la part del centre obtenint finalment seqüències de 16 imatges de 112x112. És molt important aquest retall doncs el model pre-entrenat espera aquesta resolució per a la primera capa.

Per a l'entrenament s'agafa un batch size de 8 per tal de no sobrepassar la memòria de les GPU's. En el cas de Chalearn al haver-hi 21526 vídeos per entrenament caldran 2690 iteracions per que tots el vídeos passin per la xarxa (època).

Pel cas del Dataset de Telepresence agafem també un batch size de 8 al ser els vídeos de longitud similar als de Chalearn. Per altra banda al haver-hi 240 vídeos per a fer el Train ens caldran 30 iteracions per a completar una època.

Procedim a realitzar el re-entrenament del model prèviament entrenat en la base de dades de UCF101 per a les bases de dades de Chalearn i Telepresence. Aquest re-entrenament es realitza carregant els pesos per la base de dades prèvia al model, i es re-entrena per la nova base de dades. Obtenint un mapa de pesos diferent adaptat a la nova base de dades.

Un cop entrenats i analitzats els resultats tant en la fase de entrenament com en la validació és procedeix a entrenar el model complet.

Com hem dit abans Caffe no té una bona implementació de les capes recurrents RNN i LSTM per lo que se'ns planteja intentar crear un capa pròpia que s'integri dins de Caffe o canviar de framework per a aquest últim pas.

La diferència principal que respecte aquest projecte i el dut a terme per P. Molchanov i el seu equip[] és que treballem en bases de dades on cada vídeo de cada gest ha estat segmentat. Això ens permet no haver de recórrer a la funció CTC doncs en el seu treball l'usa per identificar els instants on no hi ha cap gest.

Així doncs, seguint els passos que va seguir l'Albert Montes en el seu treball[4] es canvia de framework.

Del framework en Caffe s'agafen els pesos dels models entrenats en cadascuna de les bases de dades. Aquests seran convertits i carregats en el mateix model però implementat en Keras.

Per aquest motiu cal reconvertir el mapejat de pesos de Caffe al que espera Keras.

Per altra banda en la implementació de Keras, es realitza l'extracció de característiques de la base de dades agafant el model C3D previ i es crea un nou Dataset amb només els vectors de característiques.

Com que la última capa del model C3D es troba a la Fully Connected extraurem per a cada sub-clip un vector de 4096 característiques.

Generarem doncs una base de dades nova a partir de l'anterior amb els vectors de característiques, el "keyname" de cada vídeo i la etiqueta corresponent. I s'hi afegirà la informació temporal. De manera que es crearan els sets de característiques que es propagaran per la capa recurrent LSTM.

Aquesta capa LSTM tindrà 512 cel·les d'estat. Que son les que guardaran la informació d'estat. I 20 time-steps és a dir memoritzarà fins a 20 passos anteriors. Aquesta capa rebrà els sets de característiques creats anteriorment.

Un cop realitzats els passos anteriors es pot procedir a entrenar la capa recursiva. Que és l'ultima part del model complet.

Una vegada ja s'ha entrenat el model complet, es procedeix a analitzar els resultats tant en la fase de entrenament com de validació del model en les dues bases de dades. Com s'hi ha afegit una capa Recursiva que explota la dimensió temporal s'espera que els resultats millorin respecte els obtinguts només amb el model C3D sol.

4. Resultats

Un cop realitzats tots els passos de manera satisfactòria es procedeix a l'anàlisi de resultats.

4.1 C3D aplicat al Dataset de ChaLearn

Un cop fet el re-entrenament del model entrenat amb UCF101 analitzem els resultats obtinguts. Per aquest anàlisis ens fixarem en l'Accuracy obtingut tant en la fase de entrenament com de validació.

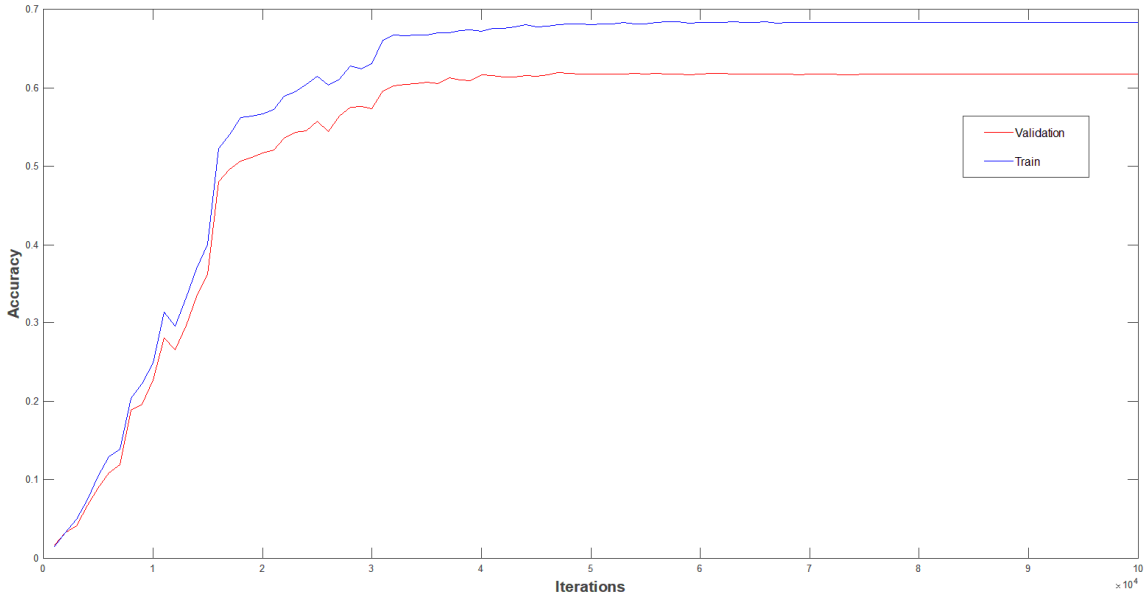

Figure 4.1: *Resultats d'aplicar el model C3D al Dataset de Chalearn*

La gràfica representa l'accuracy obtingut en cada iteració per la fase de entrenament (Blau) com de Validació (Vermell).

Veiem com el resultat és esperat. És a dir la xarxa generalitza bé malgrat no és capaç de adaptar-se del tot a la base de dades. Per comprovar aquest resultat mirem quin accuracy obtenim si testegem el model amb les dades de Test:

Model	Accuracy
C3D RGB-D Chalearn	0.616042

Table 4.1: *Test Accuracy C3D on Chalearn.*

Per tant observant els resultats obtinguts per altres projectes des de la pàgina web de Chalearn [8] podem concloure com s'han millorat els resultats.

User	Accuracy
FLiXT	0.569
AMRL	0.5557
XDETVP-TRIMPS	0.5093
ICT NHCI	0.468
XJTUfx	0.4392
TARDIS	0.4015
NTUST	0.2033

Taula 4.2: *Chalearn Challenge Results*

Comparar els resultats obtinguts pel treball de P. Molchanov [], és fa complicat doncs malgrat que van fer servir una base de dades de Chalearn preparada per al reconeixement gestual, ells van fer servir la de 2014, que contenia 14000 vídeos amb 20 gestos.

User	**Accuracy**
P. Molchanov - NVIDIA(R3DCNN)	0.966
LIRIS	0.849987
CraSPN	0.833904
JY	0.826799
CUHK-SWJTU	0.791933
Lpigou	0.788804
Stevenwudi	0.78731
Ismar	0.746631

Taula 4.3: *Resultats del Challenge de Chalearn 2014*

4.2 C3D aplicat al Dataset de Telepresence

Un cop vist que el model dona bons resultats. Procedim a veure el comportament en la base de dades de Telepresence.

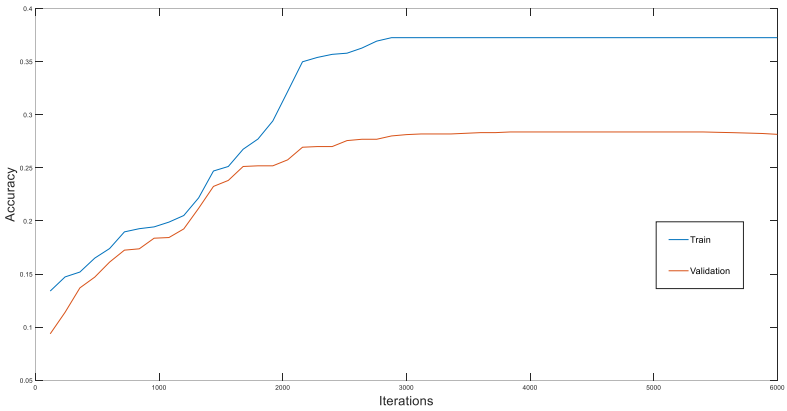

Figure 4.2: *Resultats d'aplicar el model C3D al dataset de Telepresence*

Veiem com els resultats són més irregulars i no s'aconsegueix un bon valor d'accuracy ni en la fase de entrenament ni de validació. Això ens fa pensar que el model per algun motiu no s'està adaptant bé a la base de dades malgrat no s'està produint un sobre-entrenament.

4.3 C3D + LSTM aplicat al Dataset de ChaLearn

Un cop canviat el framework i adaptat el mapejat de pesos a la framework de Keras, procedim a fer l'entrenament del model complet. Obtenint:

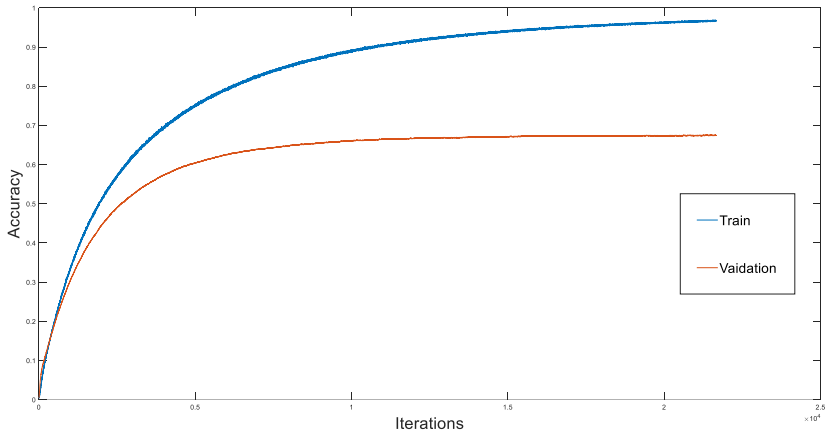

Figure 4.3: *Resultats d'accuracy del model C3DRNN aplicat a la base de dades de Chalearn*

Veiem com s'ha millorat respecte a fer servir el model C3D sol. S'ha passat de tenir un accuracy en validació del 0.61 a 0.67. Segueix sent doncs un resultat acceptable comparant amb lo obtingut per altres projectes. Si ens fixem el model s'adapta molt bé a les dades de entrenament però mai s'arriba a adaptar del a les de validació.

Anem a veure ara la gràfica de Loss per a les dues fases:

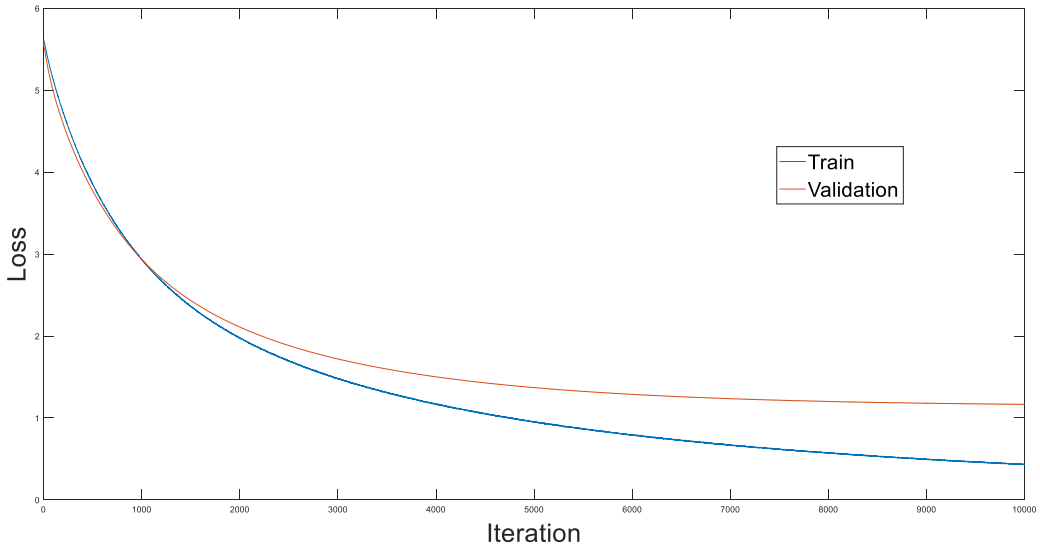

Figure 4.4: *Gràfiques del Loss al aplicar el model C3DRNN al dataset de Chalearn.*

Com podem veure la caiguda dels valors de Loss és molt semblant tant per la fase de entrenament com validació, el que ens fa veure que en cap moment s'està realitzant un sobre-entrenament.

4.4 C3D + LSTM aplicat a la Dataset de Telepresence

Finalment apliquem tot el model sobre la base de dades de Telepresence. Obtenint els següent resultats.

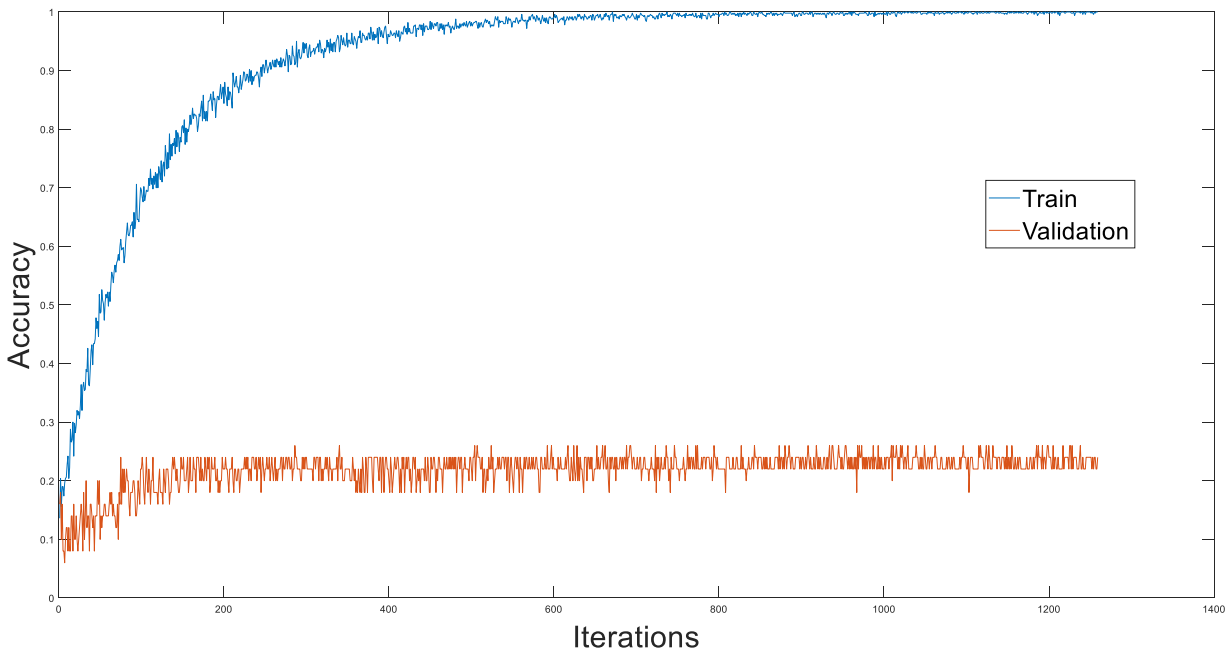

Figura 4.5: *Gràfiques d'accuracy al aplicar el model C3DRNN al dataset de Telepresence.*

Observant el gràfic de l'accuracy podem veure com el model s'adapta molt bé a les dades d'entrenament però casi gens en les de validació.

Analitzem ara les corbes de Loss.

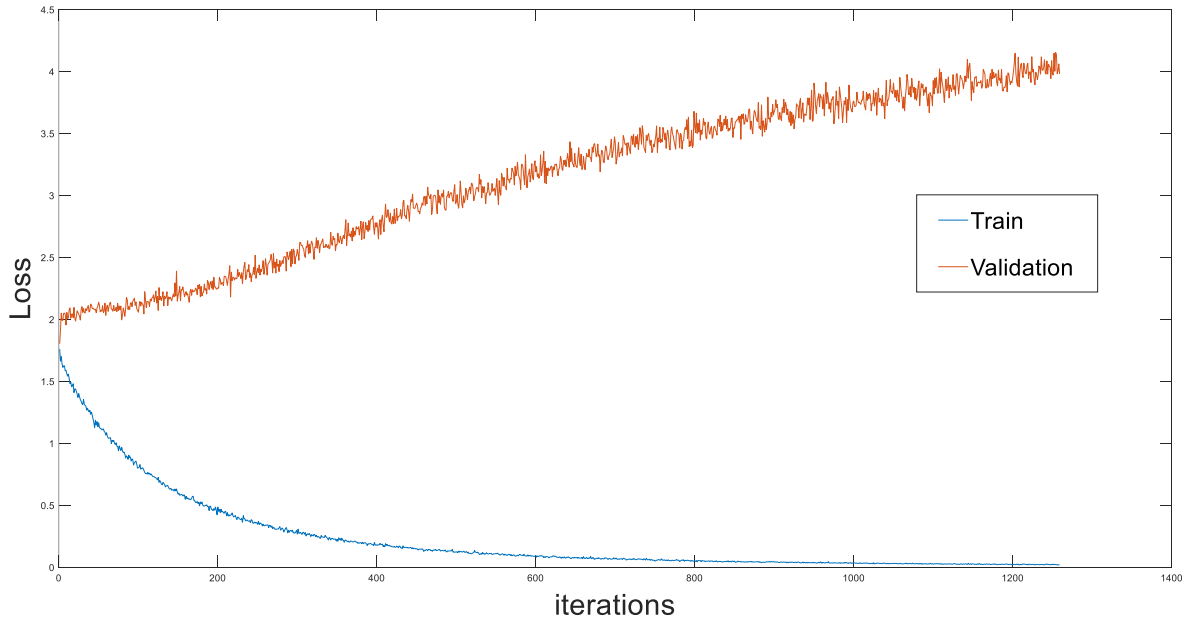

Figure 4.6: *Gràfiques del Loss al aplicar el model C3DRNN al dataset de Telepresence*

Podem observar com per validació el Loss rate no decau, al contrari s'incrementa. Això ens pot fer pensar que s'està sobre entrenant. Això pot ser degut a que al ser la base de dades petita, tot i haver realitzat un algorisme de Validació Creuada, no s'ha aconseguit que el model generalitzi prou.

Finalment i per compara el nostre model i el del treball de P. Molchanov cal esmentar els resultats que obtingueren al aplicar el seu model a la base de dades VIVA.

HOG+HOG2 [27]	color + depth	36.9%
Two-stream CNNs [34]	color + opt flow	65.6%
iDT [39]	color + opt flow	73.4%
RD3CNN	all	83.8%

Table 2.2: *Resultats en el Dataset de Viva.*

Com es pot observar obtingueren millors resultats que altres projectes, ara ve per a aconseguir-ho van fer una extracció de característiques prèvia com és el flux òptic dels vídeos. En aquest projecte no s'ha volgut obtenir del flux òptic doncs com hem dit abans es vol fer servir el deep learning per evitar haver de fer un pre-processament.

5. Costos

Tot aquest projecte s'ha realitzat mitjançant les enes que proporciona el Departament de Processament de Imatge de la UPC. Aquestes eines han sigut l'accés al servidor amb les GPU's per al processament dels vídeos, com el suport tècnic rebut. També caldrà tenir en compte les hores dedicades per el tutor Javier Ruiz. Com en qualsevol projecte d'aquest estil gran part de cost ve donat pels salaris dels investigadors i professionals implicats. El projecte ha durat 32 setmanes.

	Qty	Cost/hour	Didcated time	Total
Assistent tècnic	1	20€/h	3h/week	1920€
Senior Engineer	1	20€/h	3h/week	1920€
Junior Engineer	1	12€/h	10h/week	3840€
Servidor	1			4000€
GPU's	1			1000€
Portàtil	1			800€
			Total	13480€

Table 5.1: *Budget table*

6. Conclusions i desenvolupament futur:

Una vegada realitzats tots els passos del projecte i analitzats els resultats podem concloure que s'ha aconseguit implementar un model preparat per el reconeixement gestual entrenat en dues bases de dades.

En la primera base de dades, la de ChaLearn 2016, hem pogut millor els resultats obtinguts per altre projectes en el challenge malgrat no obtenir els resultats tant bons del projecte de P.Molchanov [2].

En quan a la base de dades de Telepresència, s'has pogut corroborar els problemes que presenta aquesta base de dades, tal com el treball de en Josep Famadas havia detectat.

Per tant malgrat aquests mal resultats per a la base de dades de Telepresència sí que podem afirmar que un model C3DRNN és un model adient per al reconeixement gestual. Afegir la capa recursiva al model C3D primari ens proporciona uns millors resultats sobretot gràcies a explotar les característiques temporals de cada vídeo.

En quan a desenvolupament futur i per seguir amb el projecte del grup de recerca caldria crear una base de dades nova o modificar la actual per tal de millorar-ne els resultats i per tant poder ser usada per al reconeixement gestual.

Bibliografia:

[1] Famadas Alsamorra, Josep. "Telos: enabling ultra-low power wireless research". Degree Thesis submitted to the Faculty of the Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona. UPC. June 2016

[2] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree and J. Kautz, "Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 4207-4215. DOI: 10.1109/CVPR.2016.456

[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks", ICCV 2015,

[4] Montes Gómez, Alberto. "Temporal activity detection in untrimmed videos with recurrent neural networks". Final Degree Thesis. UPC. June 2016.

[5] Jia, Yangqing and Shelhamer, Evan and Donahue, Jeff and Karayev, Sergey and Long, Jonathan and Girshick, Ross and Guadarrama, Sergio and Darrell, Trevor. "Caffe: Convolutional Architecture for Fast Feature Embedding". 20 June 2014.

[6] Chollet, François and others. Keras. https://keras.io 2015

[7] Li, Fei-Fei Li. Johnson , Justin. Yeung Serena. "Lecture 8: Deep Learning Software" CS231n". April 27, 2017.

[8] Jun Wan, Yibing Zhao, Shuai Zhou, Isabelle Guyon, Sergio Escalera, "ChaLearn Looking at People RGB-D Isolated and Continuous Datasets for Gesture Recognition", CVPRW, CVPR, 2016. http://chalearnlap.cvc.uab.es

Apèndixs:

Vist l'estat de l'art del reconeixement gestual és convenient establir uns conceptes bàsics que ens permetin entendre tots els passos que es vagin seguint.

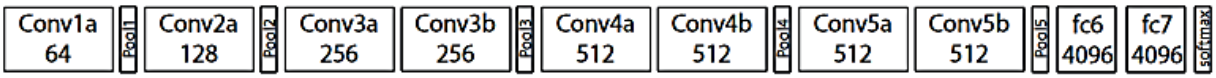

Figura A.1: *Model C3D*

Primer de tot s'explicarà el Model C3D:

Com hem dit abans el model C3D està format per 8 capes de convolució, 5 Max-Pooling, i 2 Fully Connected, i finalment una Softmax.

Es precedeix doncs a explicar aquestes diferents capes.

Convolutional layers: Son capes que donat un filtre de unes certes dimensions realitza la convolució amb l'entrada. Aquest filtre també anomenat Kernel tindrà dues dimensions en cas de les 2D-Conv, però en el cas que ens ocupa tindrà 3 dimensions. I per tant aquest convolució es realitza en les dues dimensions espacials com la temporal. És per això que en la primera capa del model C3D espera com a entrada una seqüencia d'imatges doncs realitzarà la convolució en les tres dimensions.

Molt important no confondre amb les capes 2D-Conv per a múltiples frames, doncs aquestes realitzen la convolució només en dues dimensions però en diferents frames donant com a sortida una imatge i no una seqüència. En canvi les 3D-Conv la realitzen en totes les 3 dimensions.

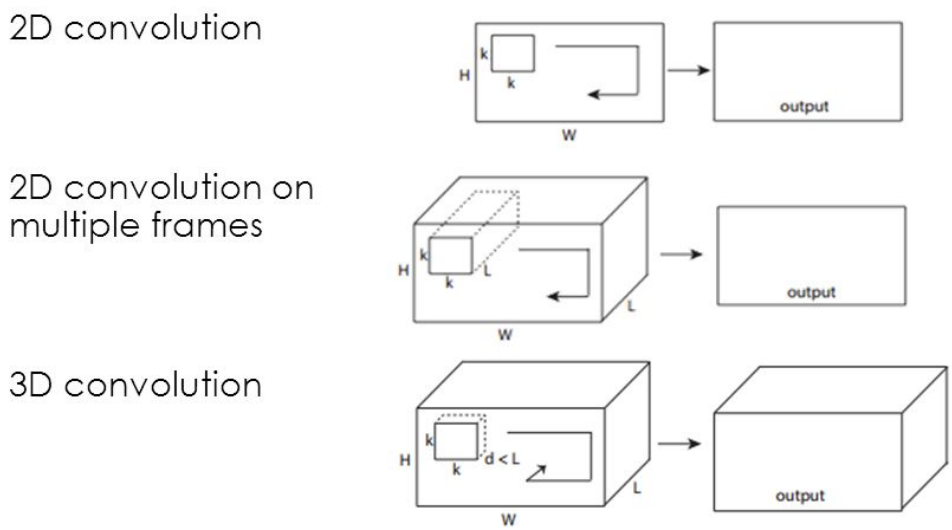

Figura A.2: *Capes convolucionals*

Un cop realitzada aquesta convolució tal com s'especifica en el Model C3D ens interessa tenir un nombre concret de sortides. Aquest nombre de sortides dependrà del nombre de filtres que s'utilitzin en cada capa.

Un cop realitzada les convolucions en tots els filtres en convindrà reduir les dimensions de les diferents sortides, per tal de reduir l'ús de memòria.

Això s'aconsegueix mitjançant una capa de MaxPooling.

Pooling Layer: aquesta ens permet reduir les dimensions de les sortides de la capa convolucional. Si ens fixem aquestes capes tenen un desplaçament major que les convolucionals això fa que es redueixin les mostres de sortida. En aquest cas concret ens trobem amb una MaxPooling Layer, que s'encarrega d'agafar de cada desplaçament les mostres més importants.

Un cop passades totes les capes convolucionals 3D ens trobem amb dues capes totalment connectades, les Fully Connected Layers.

Fully Connected Layers: Aquestes capes s'encarreguen de a partir de les característiques extretes en les capes prèvies en un vector de probabilitats.

Finalment ens trobem amb una capa final softmax.

Softmax Layer: Ens permet determinar les classes obtingudes respecte les classes de la base de dades. Obtenint així una classificació amb la que poder calcular mètriques com l'Accuracy o el Loss. És a dir mètriques que ens permetin mesurar el grau d'encert de la xarxa.

Glossari

Deep Learning: Aprenentatge profund.

Layers: Capes que formen part de la xarxa.

Accuracy: Precisió. Valor d'encert que té la xarxa a l'hora de classificar dades en classes.

Datasets. Bases de dades preparades per a ser entrenades. Moltes d'elles ja venen segmentades en Sub-sets.

Sub-sets: Segmentació dels datasets en tres grups, Entrenament, Validació i Test (Moltes vegades es prescindeix de Test).

C3D: Model RNN amb convolucions en 3 dimensions.

Sub-clip: Fragment d'un clip de vídeo.

LSTM: Long Short Term Memory, Capa recursiva que explota la dimensió temporal entre sub-clips.

Frameworks: Entorn de software amb el que es realitza una implementació.

Finetunning: Re-entrenament en una base de dades nova d'un model entrenat en una base de dades prèvia.

Cross Validation: Validació creuada. Procediment pel qual es realitza una segmentació diferent per a cadascuna de les iteracions. Fent que un model sigui entrenat i validat per totes les dades. El que ens permet poder realitzar un bon anàlisi en bases de dades petites.

Data agumentation: Procés pel qual mitjançant transformacions en les dades d'entrada s'augmenten les dimensions de la base de dades.

MaxPooling: Capa que ens permet agafar les característiques més significatives permetent reduir el nombre de característiques.

Softmax. Funció que ens permet assignar un vector de probabilitats a les classes de la base de dades.